[주제 2]

분할표 분석
-범주형 자료에 대한 비교 분석

2018년 1학기

집단 비교 방법의 이해

[분석 방법의 기본 개념]

1. 비율에 대한 가설검정
2. 교차분석
   1) 카이제곱검정
   2) 대응 자료의 집단 비교 : 맥니마검정
   3) 세 집단 이상의 비교 : 코크란-맨텔-핸젤 검정
3. 연관성 측도
4. 진단법평가

두 변수이상 관계를 분석하는 방법 소개

❖ 두 변수의 상호연관관계

자료 수준		변수 2	
		질적	계량
변수 1	질적	분할표/교차분석	(상관/연관성분석)
	계량	(상관/연관성분석)	상관분석

❖ 두 변수의 인과(종속)관계

자료 수준		종속 변수	
		질적	계량
독립 변수	질적	로그선형모형	분산분석
	계량	로지스틱회귀분석 일반화 로짓분석 판별분석 등	회귀분석

범주형자료분석의 기본 개념

- 범주형 자료의 표현과 통계량
  - 빈도(frequency) : 일반적인 범주형 자료의 표현방법
    - 빈도표(frequency table)
    - 분할표(contingency table) : 2차원 이상 자료의 표현
  - 비율(proportion) : 범주형 자료의 요약값, 대규모 자료분석에 주로 이용
    - 주로 계량분석기법으로 분석
- 범주형 자료 분석의 목표와 분석 방법
  - 집단 (차이) 비교 : 비율 차이 검정, 분포 동일성 검정
  - 변수 관계 : 독립성 검정
  - 모형 분석(일종의 인과관계) : 로짓분석, 로지스틱회귀분석 로그선형 모형

1.비율에 대한 가설검정

One Proportion

- Definition
  - the observation

$$y_i = \begin{cases} 1, & \text{조건 만족} \\ 0, & \text{o.w.} \end{cases}$$

    - Distribution of y_i

$$y_i \sim Bernouli\ (p)$$

  - Proportion

$$\hat{p} = \frac{1}{n}\sum y_i = \frac{r}{n}, \quad \text{where} \quad r = \sum y_i$$

  - Distribution of $\hat{p}$

$$\hat{p} \xrightarrow{\text{large } n \text{ and } (np>5 \text{ and } n(1\text{-}p)>5)} N\left(p, {}^{p(1-p)}\!\Big/_{n}\right)$$

- Test statistic for the null hypothesis $H_0 : p = p_0$

$$z = \frac{\hat{p} - p_0}{se(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

  - **Continuity correction** : to make the continuous data and to reduce the difference between the observation and expected proportions

$$z_C = \frac{|\hat{p} - p_0| - \frac{1}{2n}}{se(\hat{p})} = \frac{|\hat{p} - p_0| - \frac{1}{2n}}{\sqrt{p_0(1 - p_0)/n}}$$

two independent proportions

- The standard error of the difference $\hat{p}_1 - \hat{p}_2$

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{V}(\hat{p}_1) + \hat{V}(\hat{p}_2)}$$

$$= \begin{cases} \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}} & \text{, not equal variance s} \\ \sqrt{\bar{p}(1-\bar{p})}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} & \text{, equal variance s or } p_1 = p_2 \end{cases}$$

where $\bar{p} = \dfrac{r_1 + r_2}{n_1 + n_2}$

- Test statistic for the null hypothesis $H_0 : p_1 = p_2$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{se(\hat{p}_1 - \hat{p}_2)} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(1/n_1 + 1/n_2 \right)}}$$

  – **Continuity correction** :

$$z_C = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2}\left(1/n_1 + 1/n_2 \right)}{\sqrt{\overline{p}(1-\overline{p})\left(1/n_1 + 1/n_2 \right)}}$$

two paired proportions

- **Assumption** : we may observe two proportions on the same subjects
  - to compare the pain relief by two different analgesics in the same subjects or matched groups
  - to compare the proportion of subjects with a particular symptom before and after treatment

before	after	frequency
yes	yes	a
yes	no	b
no	yes	c
no	no	d

- Computation of the proportions
  - the proportions

$$\hat{p}_1 = \frac{a+b}{n} \quad \text{and} \quad \hat{p}_2 = \frac{a+c}{n}$$

  - the difference of proportions

$$\hat{p}_1 - \hat{p}_2 = \frac{a+b}{n} - \frac{a+c}{n} = \frac{b-c}{n}$$

- The standard error of the difference $\hat{p}_1 - \hat{p}_2$

$$se\left(\hat{p}_1 - \hat{p}_2\right) = \begin{cases} \dfrac{1}{n}\sqrt{b + c - \dfrac{(b-c)^2}{n}} & , \quad \text{no equal var.' s} \\ \dfrac{1}{n}\sqrt{b+c} & , \quad \text{equal var.' s or } p_1 = p_2 \end{cases}$$

- Test statistic for the null hypothesis $H_0 : p_1 = p_2$

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{se(\hat{p}_1 - \hat{p}_2)} = \frac{(b-c)/n}{\sqrt{b+c}\big/n} = \frac{(b-c)}{\sqrt{b+c}}$$

  - **Continuity correction** :

$$z_C = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2}\left(1\big/n_1 + 1\big/n_2\right)}{se\left(\hat{p}_1 - \hat{p}_2\right)}$$

$$= \frac{\frac{1}{n}|b-c| - \frac{1}{2}\left(1\big/n + 1\big/n\right)}{\sqrt{b+c}\,/\,n} = \frac{|b-c|-1}{\sqrt{b+c}}$$

2-1.교차분석

– 카이제곱검정

분할표 자료의 표현 방법

- 질적(혹은 범주형)자료는 비율로 요약
- 독립성과 분포 동일성 자료의 표현 방법

독립성		열 변수			합계
		1	⋯	c	
행변수	1	n_{11}	⋯	n_{1c}	$n_{1\bullet}$
	⋯	⋯	⋯	⋯	⋯
	r	n_{r1}	⋯	n_{rc}	$n_{r\bullet}$
합계		$n_{\bullet 1}$	⋯	$n_{\bullet c}$	n

확률변수

고정

분포 동일성		열 변수			합계
		1	⋯	c	
행변수	1	n_{11}	⋯	n_{1c}	$n_{1\bullet}$
	⋯	⋯	⋯	⋯	⋯
	r	n_{r1}	⋯	n_{rc}	$n_{r\bullet}$

고정

사례 : 독립적인 두 집단의 분석

- 남자 중 지난 해 병원을 방문한 비율과 여자 중 지난 해 병원을 방문 비율에 차이가 있을까?
  - 비율 차이 검정(혹은 z-검정)
  - 카이제곱(χ^2 혹은 X^2)검정 : 동질성(분포 동일성) 검정

- 성별과 병원방문 여부 간에 연관관계가 있을 까?
  - 카이제곱 검정 : 독립성 검정

연관성 검정(카이제곱 검정)의 가설

- 분포의 동일성(혹은 동질성) 검정
  - 귀무가설 : 성별에 따른 (남녀의) 병원 방문 여부(비율)의 분포는 동일하다(의미 : 남자의 병원 방문비율과 여자의 병원 방문비율은 차이가 없다 – 비율차이 검정 관점).
  - 대립가설 : 남녀의 병원 방문비율은 동일하지 않다(의미 : 남자의 병원 방문 비율과 여자의 병원 방문 비율은 다르다).

- 독립성 검정
  - 귀무가설 : 성별과 병원 방문여부는 관계가 없다.
  - 대립가설: 성별과 병원 방문여부는 관계가 있다.

독립성 검정을 위한 기본 개념의 이해

- 독립사건의 확률 : 서로 독립적인 A와 B의 사건이 우연히 동시에 일어날 확률은?

$$P(A \text{ and } B) = P(A) \cdot P(B) = p_{ij} = p_{i} \cdot p_{j} = \frac{n_{i\bullet}}{n} \times \frac{n_{\bullet j}}{n}$$

  - 예 : 여자(F)가 병원을 우연히 방문(V)할 확률?

    P(F and V)=P(F) * P(V)=(F/n)*(V/n)

    =(276/408)*(210/408)=0.348

- 독립사건의 기대빈도 계산
  - (i행 j열) 범주의 기대 빈도 계산 $n \cdot p_{ij} = n \cdot p_{i} \cdot p_{j} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$
  - 예 : 여자가 병원을 우연히 방문할 것으로 예상되는 사람의 수(기대빈도)?

$$n \cdot p_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} = \frac{276 \cdot 210}{408} = 142.1$$

카이제곱 검정

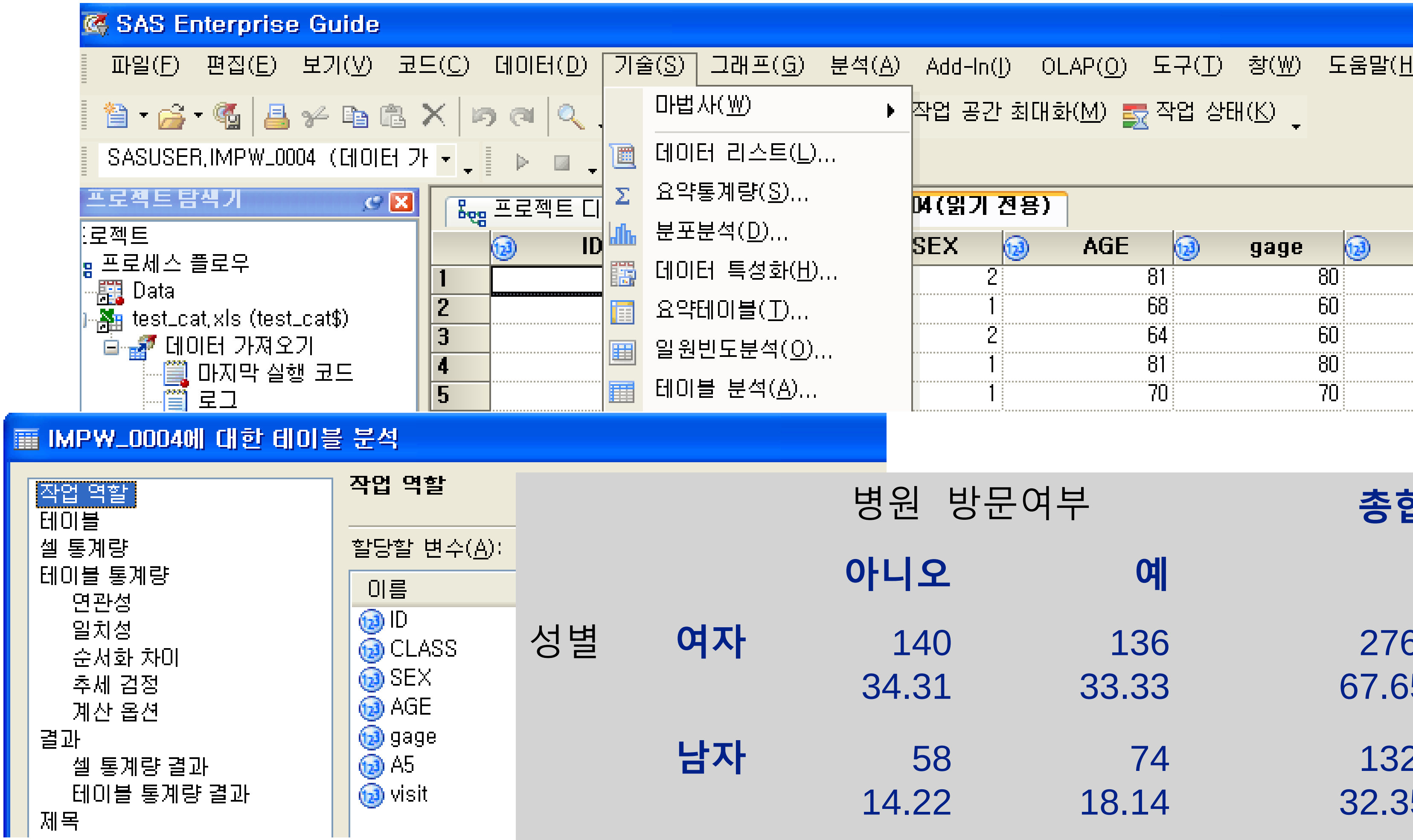
- 검정통계량은 관측빈도와 기대빈도의 차이에 기초하여 계산

$$\chi^2 = \sum_{\text{모든 칸}} \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}}$$

- 귀무가설이 옳다는 가정하에서 검정통계량은 카이제곱분포 : 자유도=(열의 수-1)*(행의 수-1)

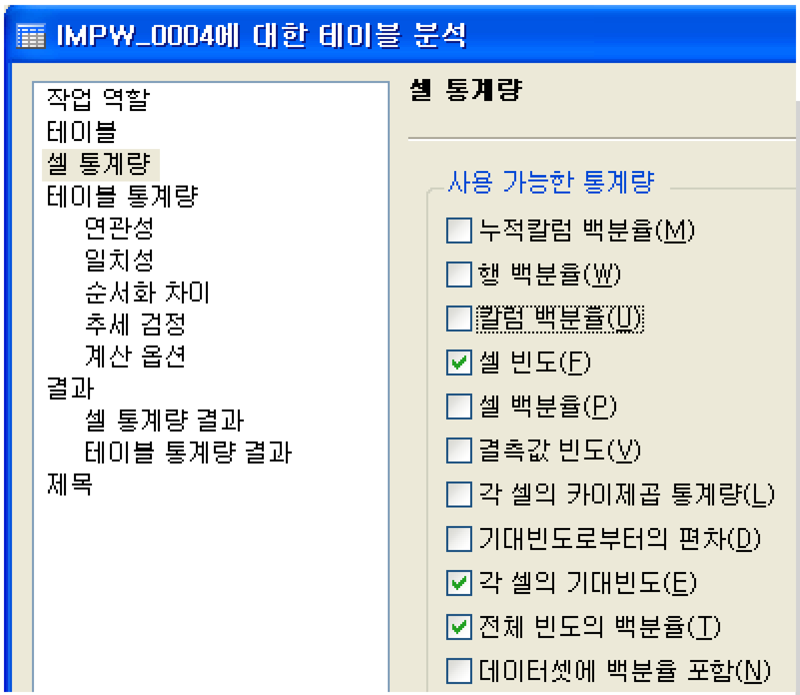
- 판단 기준
  - 검정통계량 값이 클수록 귀무가설을 기각할 확률이 높아지므로 두 변수 사이가 독립이 아니라는 의미임. 즉, 서로 연관관계를 가지고 있다고 해석
  - 카이제곱 검정으로 유의성 판단

Contingency Table : SAS



		병원 방문여부		총합
		아니오	예	
성별	여자	140 34.31	136 33.33	276 67.65
	남자	58 14.22	74 18.14	132 32.35
총합		198 48.53	210 51.47	408 100.00

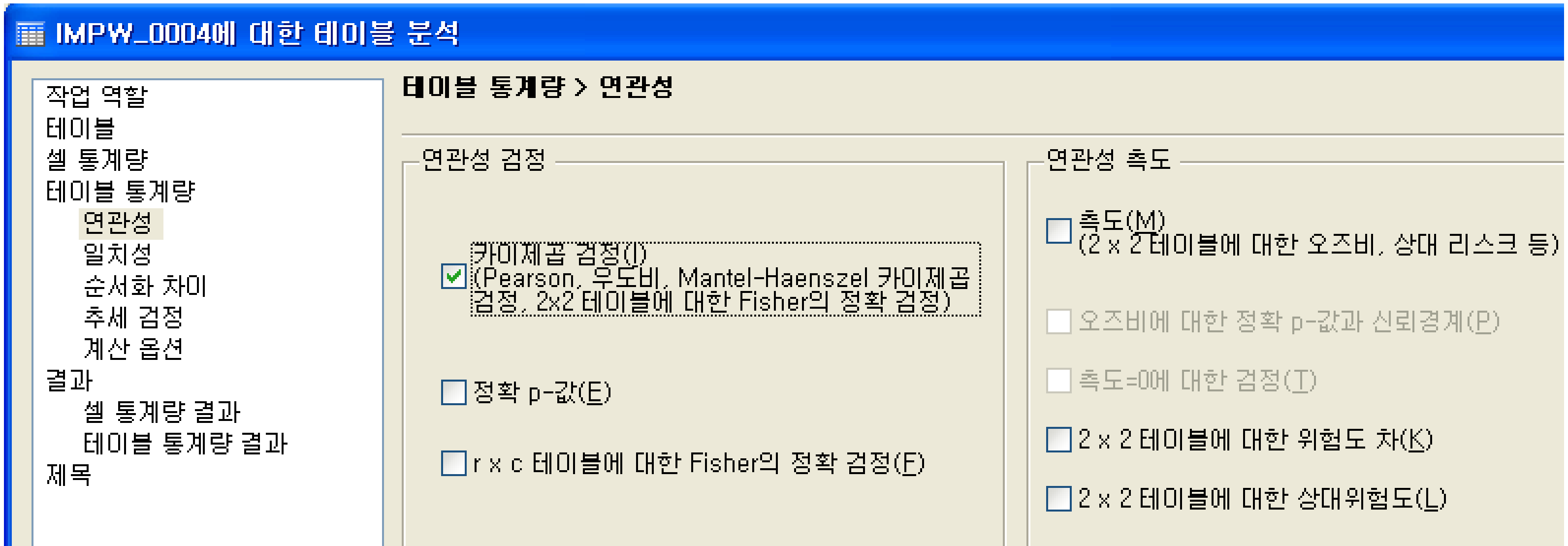
관측빈도와 기대빈도 : SAS



		병원 방문여부		총합
		아니오	예	
성별	여자	140 133.94 34.31	136 142.06 33.33	276 67.65
	남자	58 64.059 14.22	74 67.941 18.14	132 32.35
총합		198 48.53	210 51.47	408 100.00

만약 기대빈도와 관측빈도의 차이가 크다면 두 변수가 서로 독립(무관)이다는 귀무가설을 의심(독립이 아니라는 의미)

카이제곱 검정 : SAS 결과 및 해석



통계량	자유도	값	확률
카이제곱	1	1.6458	0.1995
우도비 카이제곱	1	1.6492	0.1991

유의확률이 0.2수준이므로 5%의 유의수준에서 귀무가설을 기각하지 못하므로 독립이라는 주장을 채택, 따라서 성별과 병원 방문여부는 서로 독립이라고 해석(두 변수 사이에 관련성이 없음을 의미함)

카이제곱 검정의 이해

- 표본수가 작을 때에는 타당도가 떨어짐
- 기대빈도가 5이하인 셀의 수가 20% 이하이고, 1 이하의 기대빈도를 가진 셀이 없어야 타당성이 인정
  - 만약 이 조건이 만족되지 않으면 Fisher's Exact test 수행
- 표본크기가 50이하인 경우이거나 이산적인 분포에 대하여 연속적인 카이제곱분포로 근사 적용
  - 2*2 table에서 Fisher's exact test, Yates correction(연속성부여) 으로 보정

$$\chi^2 = \sum_{\text{모든 칸}} \frac{(\text{관측빈도} - \text{기대빈도} - 1/2)^2}{\text{기대빈도}}$$

- 분포동일성 검정은 가설(과 해석)만 다를 뿐 검정의 모든 과정은 독립성과 동일한 방법으로 검정

Fisher의 정확 검정법

- 카이제곱검정은 근사검정이므로 기대도수가 작은 경우 카이제곱검정통계량의 분모값이 작아져 통계량을 과대 계산하게 되므로 부적합 검정으로 인지됨

- 기대도수가 5이하인 셀이 20% 초과하는 경우 카이제곱 검정대신 피셔의 정확 검정법(fisher's exact test)을 사용하는 것이 바람직
  - 기본 가정 : 행의 합과 열의 합이 고정(fixed)
  - p 값의 계산

$$p-\text{값} = \text{관측된 분할표가 나올 확률} + \text{관측된 분할표보다 더 극단적인 분할표가 나올 확률}$$

예제

- 예제 : 10명 환자들에게 처치 1과 2를 받게 한 후 나타난 반응 결과이다. 처치와 반응 결과간에 연관성이 존재하는 지 판단하라.

	반응 결과		합
	반응	무반응	
처치 1	1	4	5
처치 2	3	2	5
합	4	6	10

H_0 : 처치와 반응은 관계가 없다

H_1 : 처치 2의 반응율이 처치 1의 반응율보다 높다

(단측 검정)

[풀이] 2x2 분할표이고, 행과 열의 합이 고정된 것으로 가정하였기에 (1,1)칸의 도수만 알면 나머지 칸의 도수는 자동으로 결정

1) 처치 1의 반응에 대한 확률(도수=1)

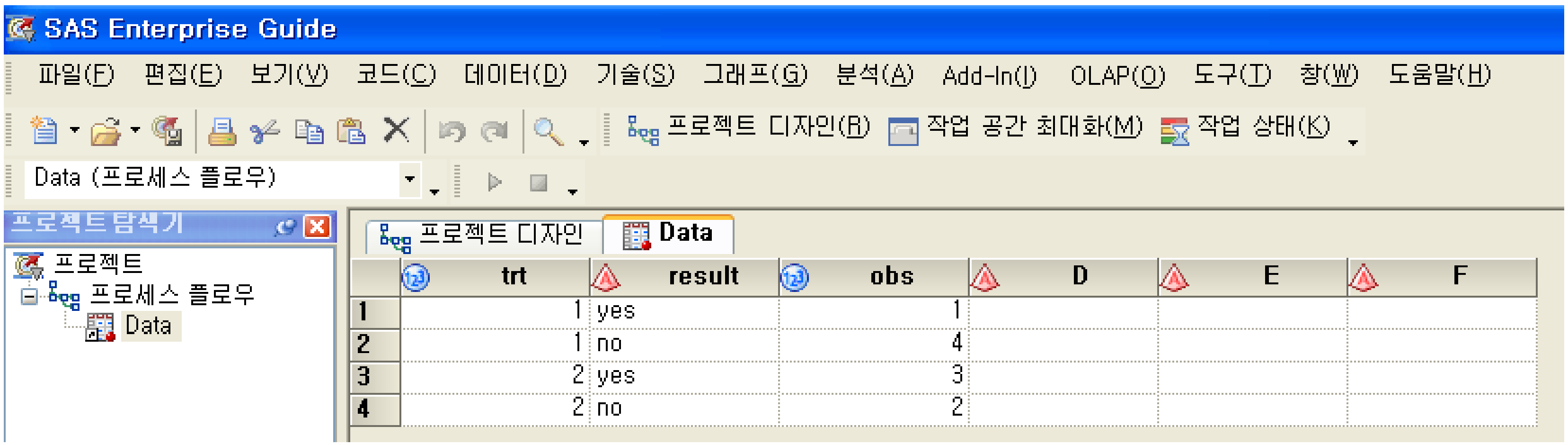
$$\Pr(n_{11}=1)=\frac{\binom{4}{1}\binom{6}{4}}{\binom{10}{5}}=0.238$$

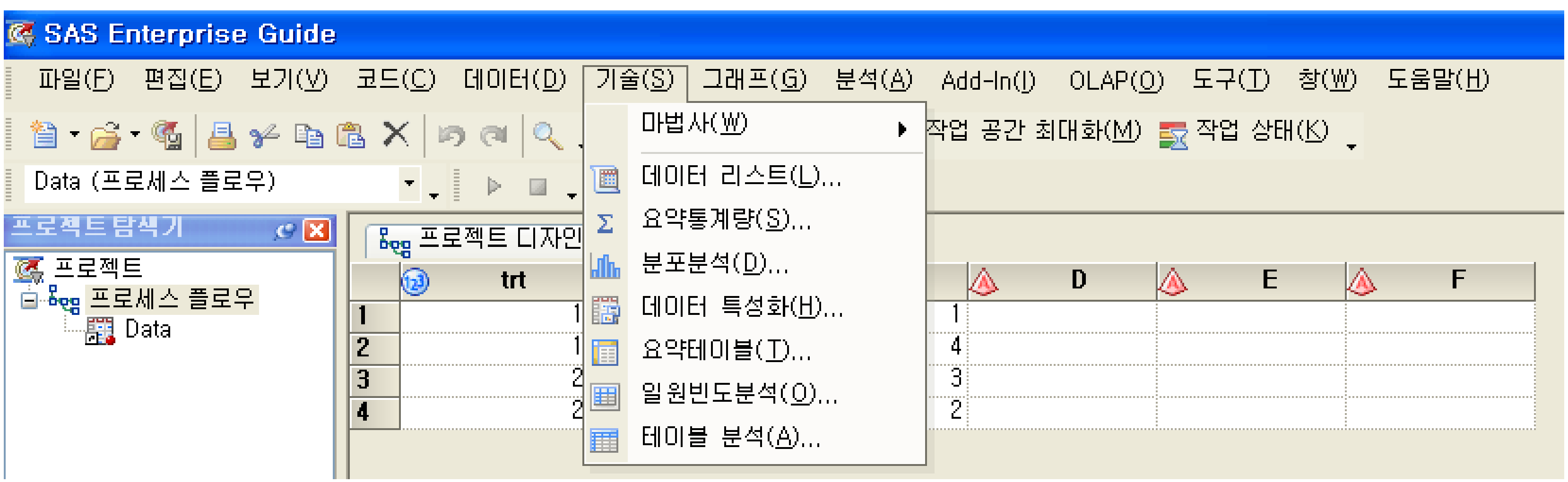
2) 처치 1의 반응에 대한 극단적인 분할표의 확률(도수=0)

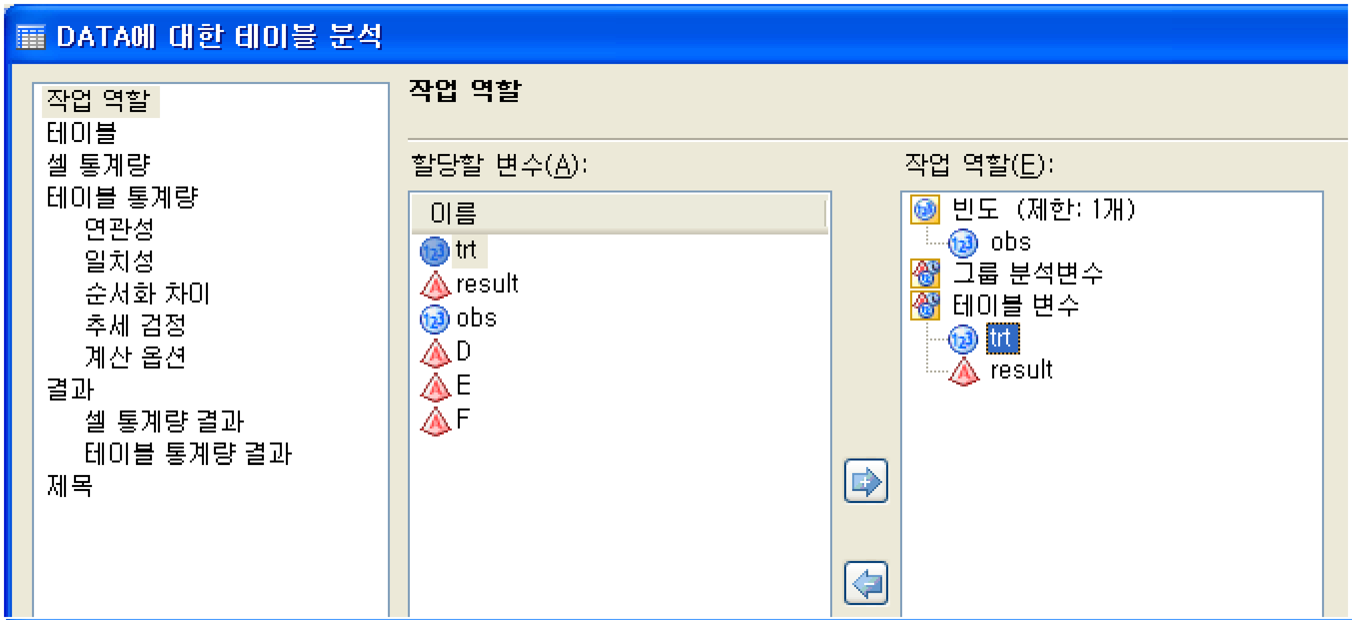
$$\Pr(n_{11}=0)=\frac{\binom{4}{0}\binom{6}{5}}{\binom{10}{5}}=0.024$$

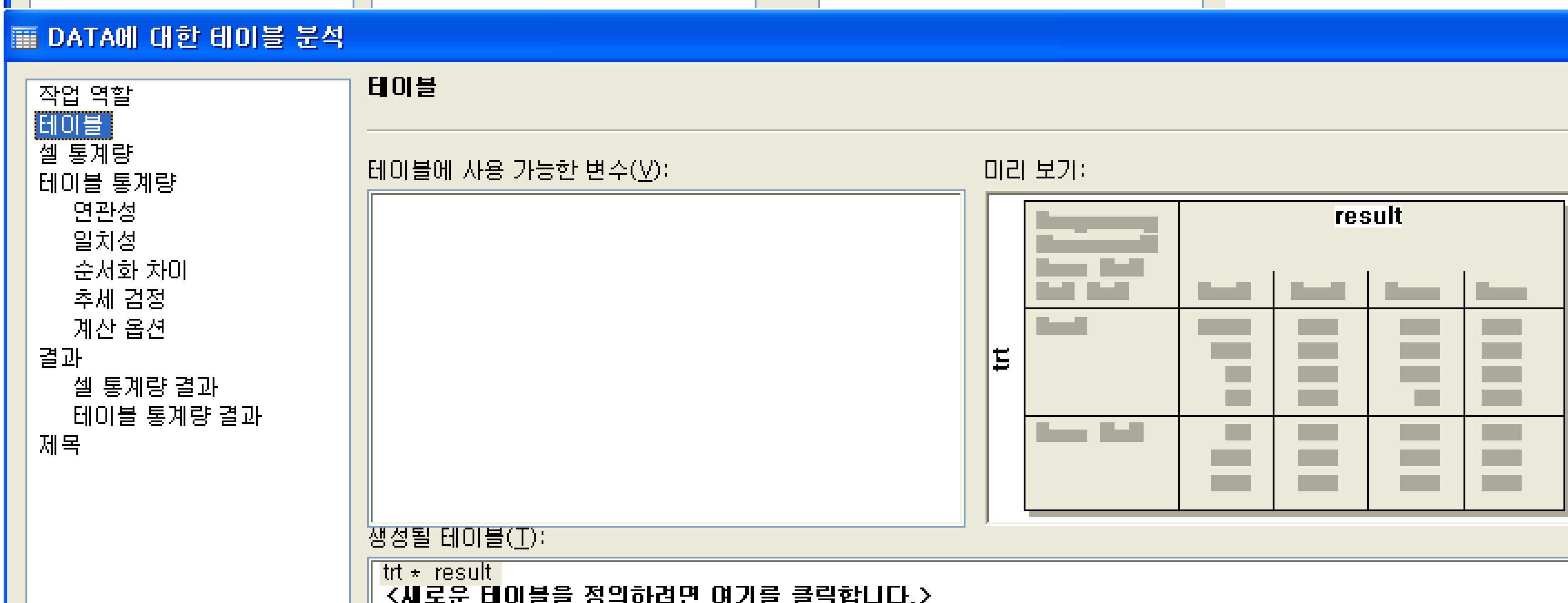
- p값=0.238+0.024=0.262이므로 유의수준 5%에서 귀무가설은 기각되지 않으므로 처치와 반응 사이에는 관계가 없다고 판단한다.

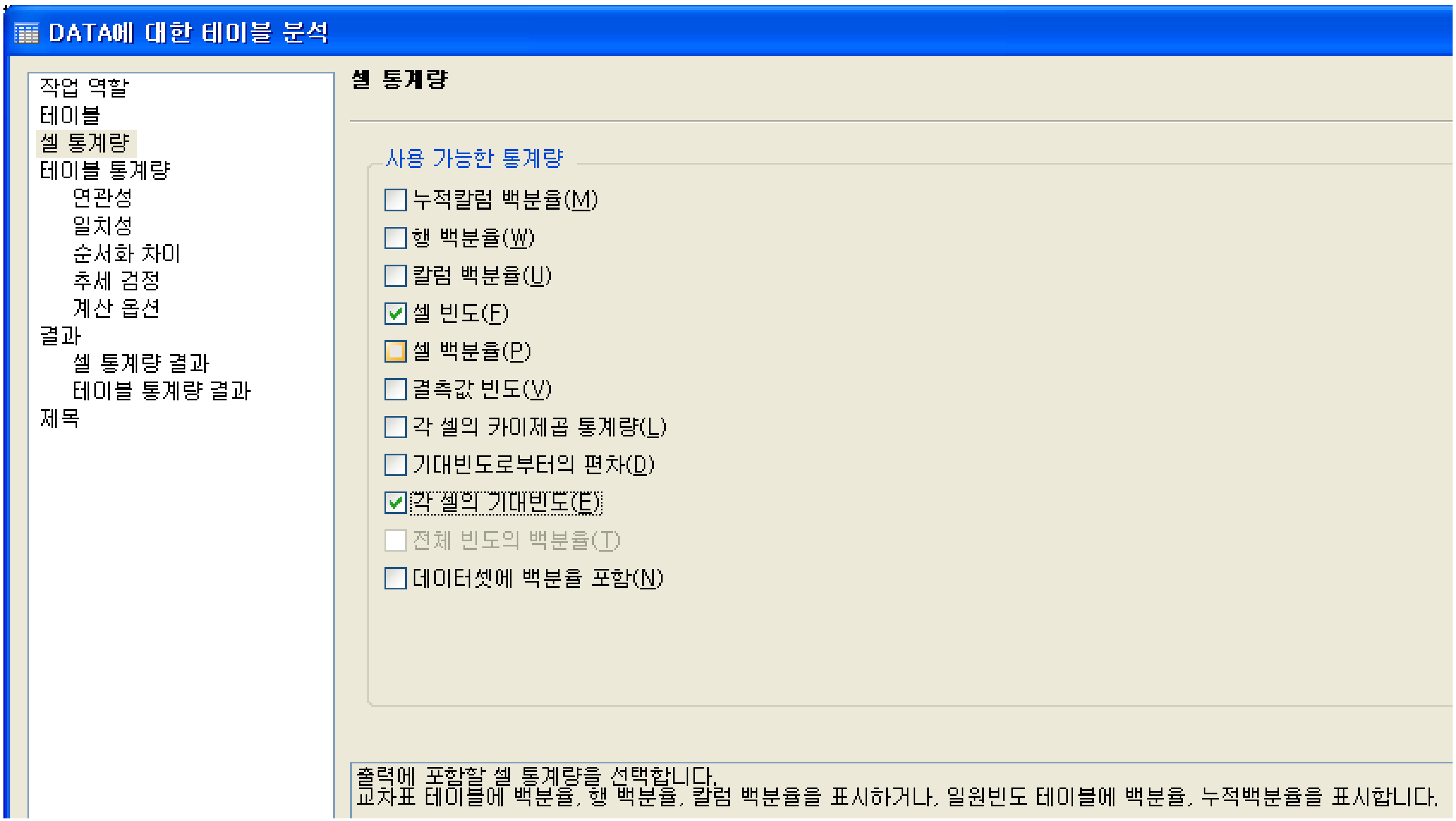
Example : Fisher's exact test

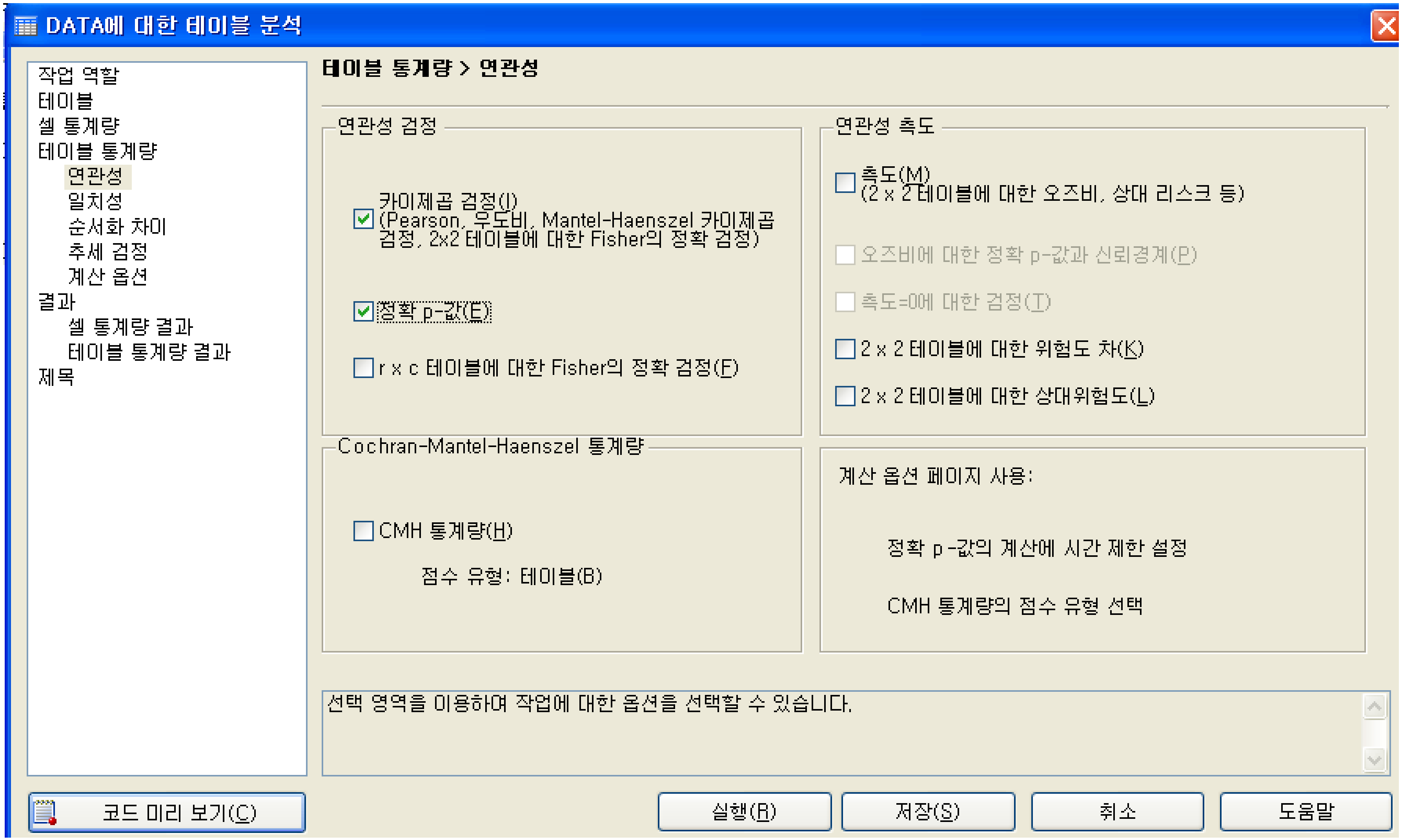





DATA에 대한 테이블 분석
작업 역할
테이블
셀 통계량
테이블 통계량
연관성
일치성
순서화 차이
추세 검정
계산 옵션
결과
셀 통계량 결과
테이블 통계량 결과
제목
작업 역할
할당할 변수(A):
이름
trt
result
obs
D
E
F
작업 역할(E):
빈도 (제한: 1개)
obs
그룹 분석변수
테이블 변수
trt
result


DATA에 대한 테이블 분석
작업 역할
테이블
셀 통계량
테이블 통계량
연관성
일치성
순서화 차이
추세 검정
계산 옵션
결과
셀 통계량 결과
테이블 통계량 결과
제목
테이블
테이블에 사용 가능한 변수(V):
미리 보기:
result
trt
생성될 테이블(T):
trt * result
<새로운 테이블을 정의하려면 여기를 클릭합니다.>


DATA에 대한 테이블 분석
작업 역할
테이블
셀 통계량
테이블 통계량
연관성
일치성
순서화 차이
추세 검정
계산 옵션
결과
셀 통계량 결과
테이블 통계량 결과
제목
셀 통계량
사용 가능한 통계량
누적칼럼 백분율(M)
행 백분율(W)
칼럼 백분율(U)
셀 빈도(F)
셀 백분율(P)
결측값 빈도(V)
각 셀의 카이제곱 통계량(L)
기대빈도로부터의 편차(D)
각 셀의 기대빈도(E)
전체 빈도의 백분율(T)
데이터셋에 백분율 포함(N)
출력에 포함할 셀 통계량을 선택합니다.
교차표 테이블에 백분율, 행 백분율, 칼럼 백분율을 표시하거나, 일원빈도 테이블에 백분율, 누적백분율을 표시합니다.


DATA에 대한 테이블 분석
작업 역할
테이블
셀 통계량
테이블 통계량
연관성
일치성
순서화 차이
추세 검정
계산 옵션
결과
셀 통계량 결과
테이블 통계량 결과
제목
테이블 통계량 > 연관성
연관성 검정
카이제곱 검정(I)
(Pearson, 우도비, Mantel-Haenszel 카이제곱 검정, 2x2 테이블에 대한 Fisher의 정확 검정)
정확 p-값(E)
r x c 테이블에 대한 Fisher의 정확 검정(F)
연관성 측도
측도(M)
(2 x 2 테이블에 대한 오즈비, 상대 리스크 등)
오즈비에 대한 정확 p-값과 신뢰경계(P)
측도=0에 대한 검정(T)
2 x 2 테이블에 대한 위험도 차(K)
2 x 2 테이블에 대한 상대위험도(L)
Cochran-Mantel-Haenszel 통계량
CMH 통계량(H)
점수 유형: 테이블(B)
계산 옵션 페이지 사용:
정확 p -값의 계산에 시간 제한 설정
CMH 통계량의 점수 유형 선택
선택 영역을 이용하여 작업에 대한 옵션을 선택할 수 있습니다.
코드 미리 보기(C)
실행(R)
저장(S)
취소
도움말

SAS Output

trt * result 테이블

trt	result		총합
	no	yes	
1	4 3	1 2	5
2	2 3	3 2	5
총합	6	4	10

통계량	자유도	값	확률
카이제곱	1	1.6667	0.1967
우도비 카이제곱	1	1.7261	0.1889
연속성 수정 카이제곱	1	0.4167	0.5186
Mantel-Haenszel 카이제곱	1	1.5000	0.2207
파이 계수		0.4082	
우발성 계수		0.3780	
크래머의 V		0.4082	

경고: 셀의 100%가 5보다 적은 기대빈도를 가지고 있습니다.
(근사)카이제곱 검정은 올바르지 않을 수 있습니다.

Pearson 카이제곱 검정	
카이제곱	1.6667
자유도	1
근사적인 Pr > ChiSq	0.1967
정확한 Pr >= ChiSq	0.5238

우도비 카이제곱 검정	
카이제곱	1.7261
자유도	1
근사적인 Pr > ChiSq	0.1889
정확한 Pr >= ChiSq	0.5238

Mantel-Haenszel 카이제곱 검정	
카이제곱	1.5000
자유도	1
근사적인 Pr > ChiSq	0.2207
정확한 Pr >= ChiSq	0.5238

Fisher의 정확 검정	
(1,1) 셀 빈도(F)	4
하단측 p값 Pr <= F	0.9762
상단측 p값 Pr >= F	0.2619
테이블 확률 (P)	0.2381
양측 p값 Pr <= P	0.5238

양측검정의 p값이 0.5238이므로 단측검정은 양측x(1/2)이므로 0.2619이 되며, 귀무가설은 기각되지 않으므로 처치와 반응사이에 연관이 없는 귀무가설을 채택

- SAS program

```
Data test_ex2;
  input  trt  result $  obs;
  cards;
  1 yes 1
  1 no 4
  2 yes 3
  2 no 2
;
Proc freq;
 Weight obs;
 Tables trt*result/exact;
Run;
```

Statistics for Table of trt by result

Statistic	DF	Value	Prob
Chi-Square	1	1.6667	0.1967
Likelihood Ratio Chi-Square	1	1.7261	0.1889
Continuity Adj. Chi-Square	1	0.4167	0.5186
Mantel-Haenszel Chi-Square	1	1.5000	0.2207
Phi Coefficient		0.4082	
Contingency Coefficient		0.3780	
Cramer's V		0.4082	

WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Cell (1,1) Frequency (F)	4
Left-sided Pr <= F	0.9762
Right-sided Pr >= F	0.2619
Table Probability (P)	0.2381
Two-sided Pr <= P	0.5238

Sample Size = 10

2-2. 맥니마 검정

– 비모수검정

2-2-2. 비모수검정: McNemar's test

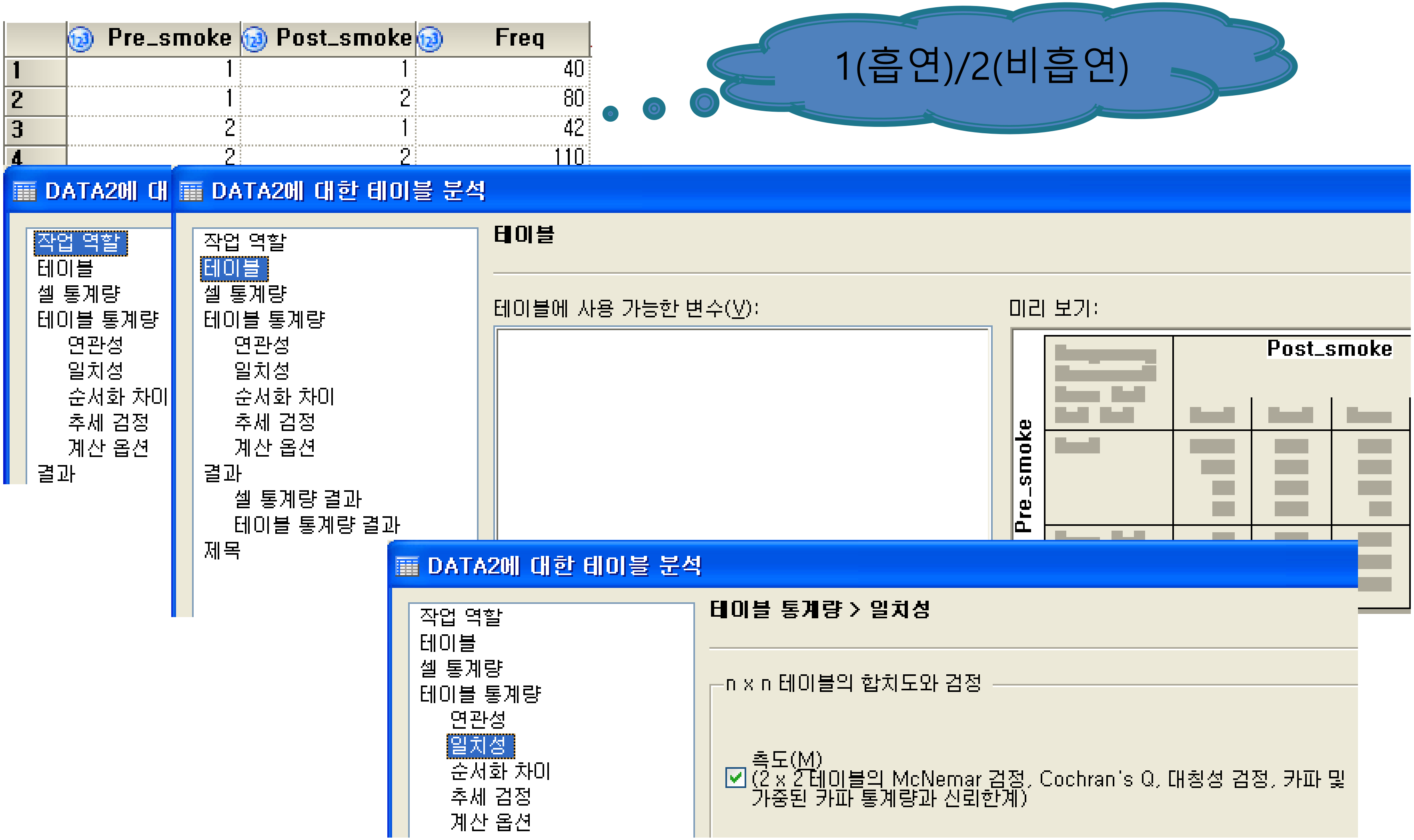
- (동일대상의) 대응자료 t-검정과 동일한 비모수적 방법
  - 예제 1 : 처리전 양성반응인 사람이 처리후 반응의 변화 여부 분석
  - 예제 2 : 금연 교육 후 1년이 지난 후, 흡연 여부를 조사하여 금연 교육의 효과 분석(동일 집단 대상)
- 연구문제: 금연교육 전의 (비)흡연율과 금연교육 후의 (비)흡연율이 동일한가? 변화가 있는 가?
  - 짝지은 표본의 범주형(빈도)자료일 때 McNemar test 실시
- 검정통계량:

$$Q_M = \frac{(\text{변화가 일어난 빈도의 차})^2}{\text{변화가 일어난 빈도의 합}} \sim \chi^2(1)$$

  - 연속적 보정 통계량(Walker, 1997)

$$Q_{M(2)} = \frac{(|\text{변화가 일어난 빈도의 차}| - 1)^2}{\text{변화가 일어난 빈도의 합}} \sim \chi^2(1)$$

SAS : McNemar's test



SAS : McNemar's test

		교육 후		총합
		흡연	비흡연	
교육 전	흡연	40 33.3	80 66.7	120 (44.1)
	비흡연	42 27.6	110 72.4	152 (55.9)
총합		82 30.1	190 69.9	272

– 교육 전 흡연율(44.1%)보다 교육 후 흡연율(30.1%)로 감소

– 교육 전의 흡연자가 교육 후 비흡연자로 된 사람은 80명, 반대의 예는 42명으로 변화.

McNemar 검정	
통계량 (S)	11.8361
자유도	1
Pr > S	0.0006

– 검정 통계량

$$Q_M = \frac{(80 - 42)^2}{80 + 42} = 11.84$$

– 해석 : 교육 전에 비해 교육 후 흡연율이 감소한 결과는 통계적으로 유의(p=0.0006)하므로 금연 교육은 효과가 있다고 해석

2-3. 코크란-맨텔-핸젤 검정

– 비모수검정

2-3. Cochran-Mantel-Haenzel test

- 둘 이상의 그룹이 있을 때 두 처리간 차이나 관계가 있는지를 검정하는 비모수적 방법으로 그룹의 효과(영향)를 제외한 두 처리간 차이를 검정
  - 독립된 k개 층이 있을 때 층의 영향을 제어한 후 처리와 결과와의 관계나 처리간 반응의 차이를 검정하는 방법
  - 예제 1 : 층화표본설계에서 얻은 자료에서 층 효과를 제어한 후 처리와 반응의 관계유무를 검정
  - 예제 2 : 병원규모나 성별 혹은 지역에 따른 차이를 제외한 후 치료제 A와 B의 반응 결과에 차이가 있는 지 살펴보고자 함

- 자료 형태

층	처치	결과		합계
		Yes	No	
1	1	n_111	n_112	n_11+
	2	n_121	n_122	n_12+
	계	n_1+1	n_1+2	n_1++
………..		………..	………..	………..
k	1	n_k11	n_k12	n_k1+
	2	n_k21	n_k22	n_k2+
	계	n_k+1	n_k+2	n_k++
합계		n_++1	n_++2	n_+++

- 가설 $H_0 : p_1 = p_2 \quad vs \quad H_1 : p_1 \neq p_2$
  - 여기서 p_1 은 처치 1의 반응율, p_2 는 처치 2의 반응율
- 검정통계량

$$Q_{CMH} = \frac{\left(\sum_{j=1}^{k} n_{j11} - \sum_{j=1}^{k} E\left[n_{j11} \mid H_0 \right] \right)^2}{\sum_{j=1}^{k} Var\left[n_{j11} \mid H_0 \right]} \sim \chi^2(1) \ \text{under} \ H_0$$

  - 여기서

$$E\left[n_{j11} \mid H_0 \right] = \frac{n_{j1+} \times n_{j+1}}{n_{j++}} \ \text{and} \ Var\left[n_{j11} \mid H_0 \right] = \frac{n_{j1+} \times n_{j2+} \times n_{j+1} \times n_{j+2}}{n_{j++}^2 \left(n_{j++} - 1 \right)}$$

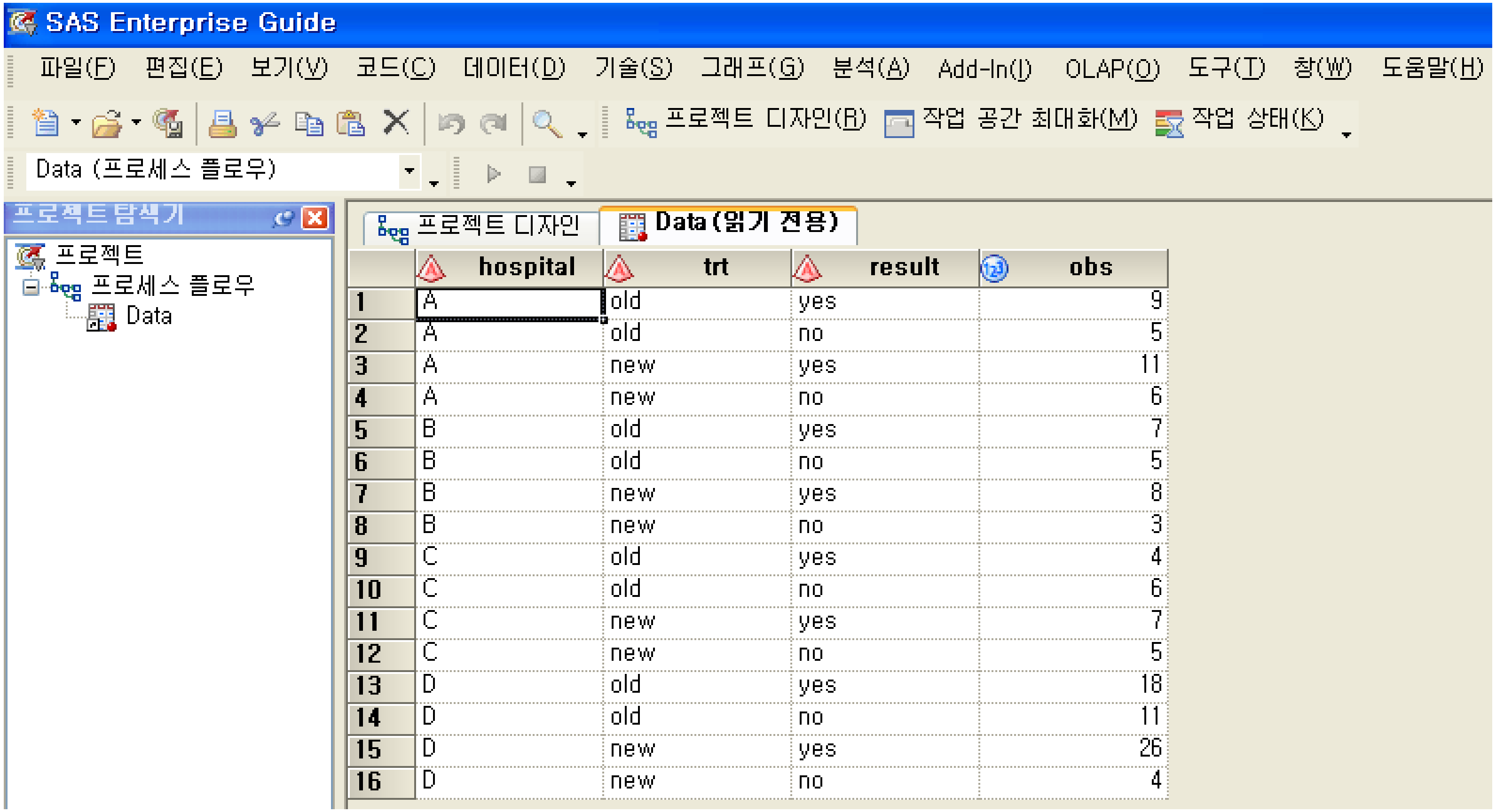
- 판단

if $\alpha > p - value$ then H_0 is rejected.

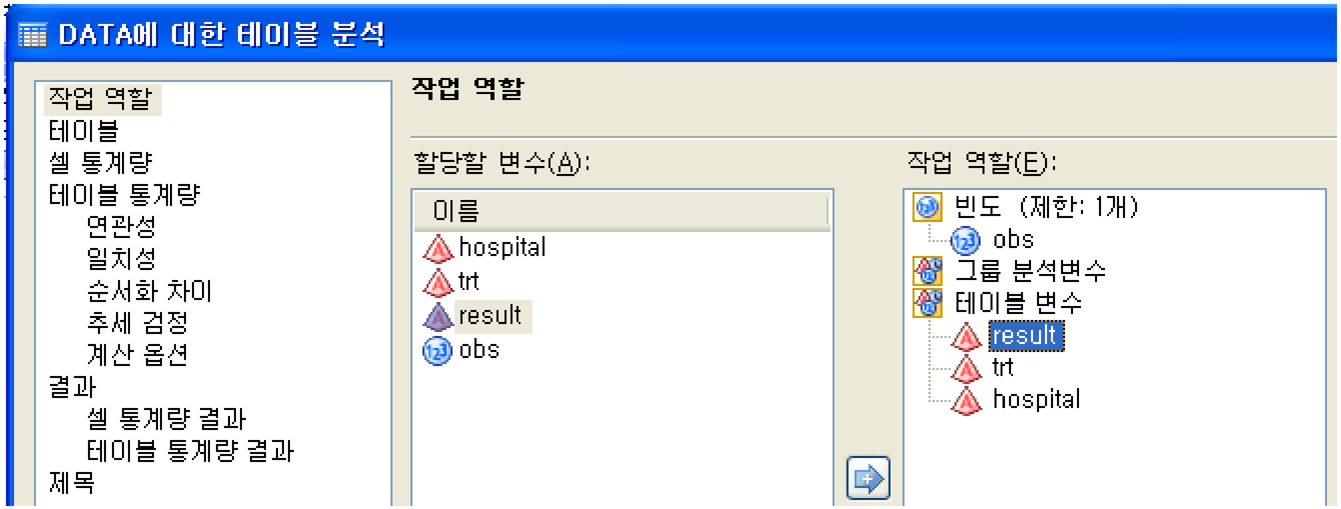
Example : SAS C-M-H test

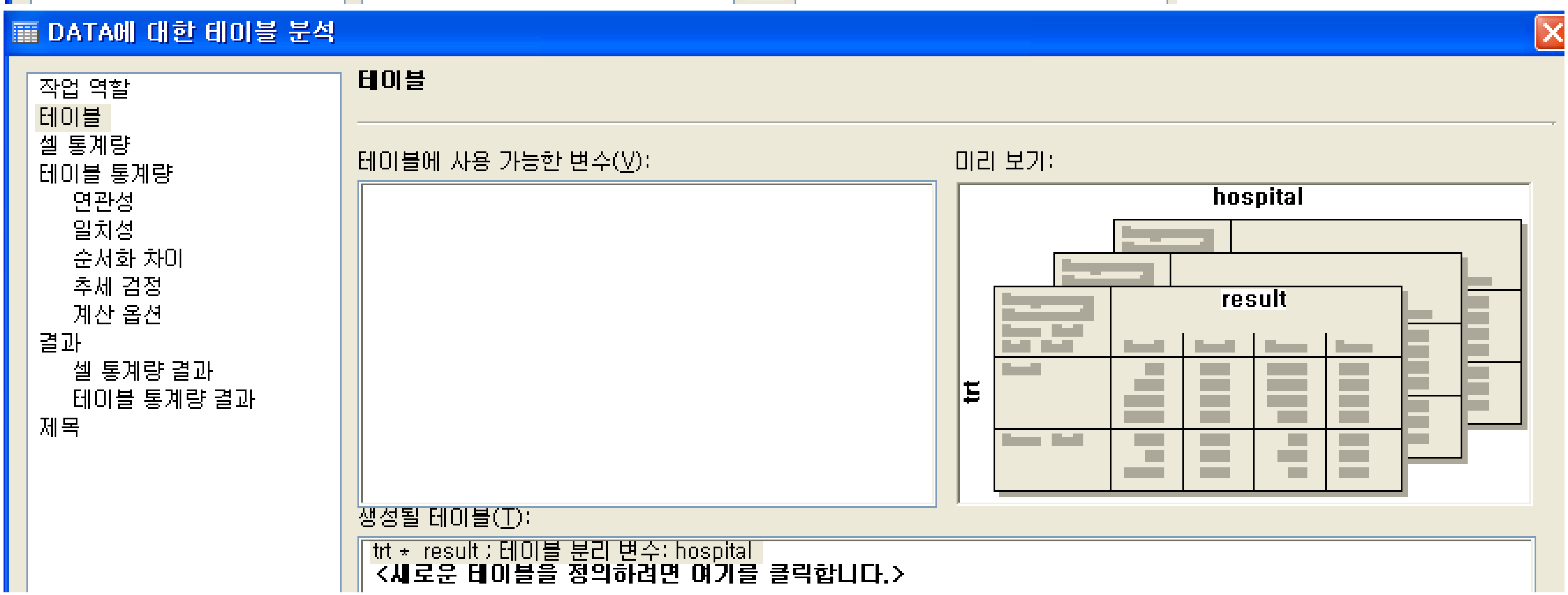
- 예제 : 호흡기 환자의 치료제에 대한 효과
  - 병원의 영향을 제거한 후 두 치료제의 호전도 비교(랜덤배분)
  - 자료

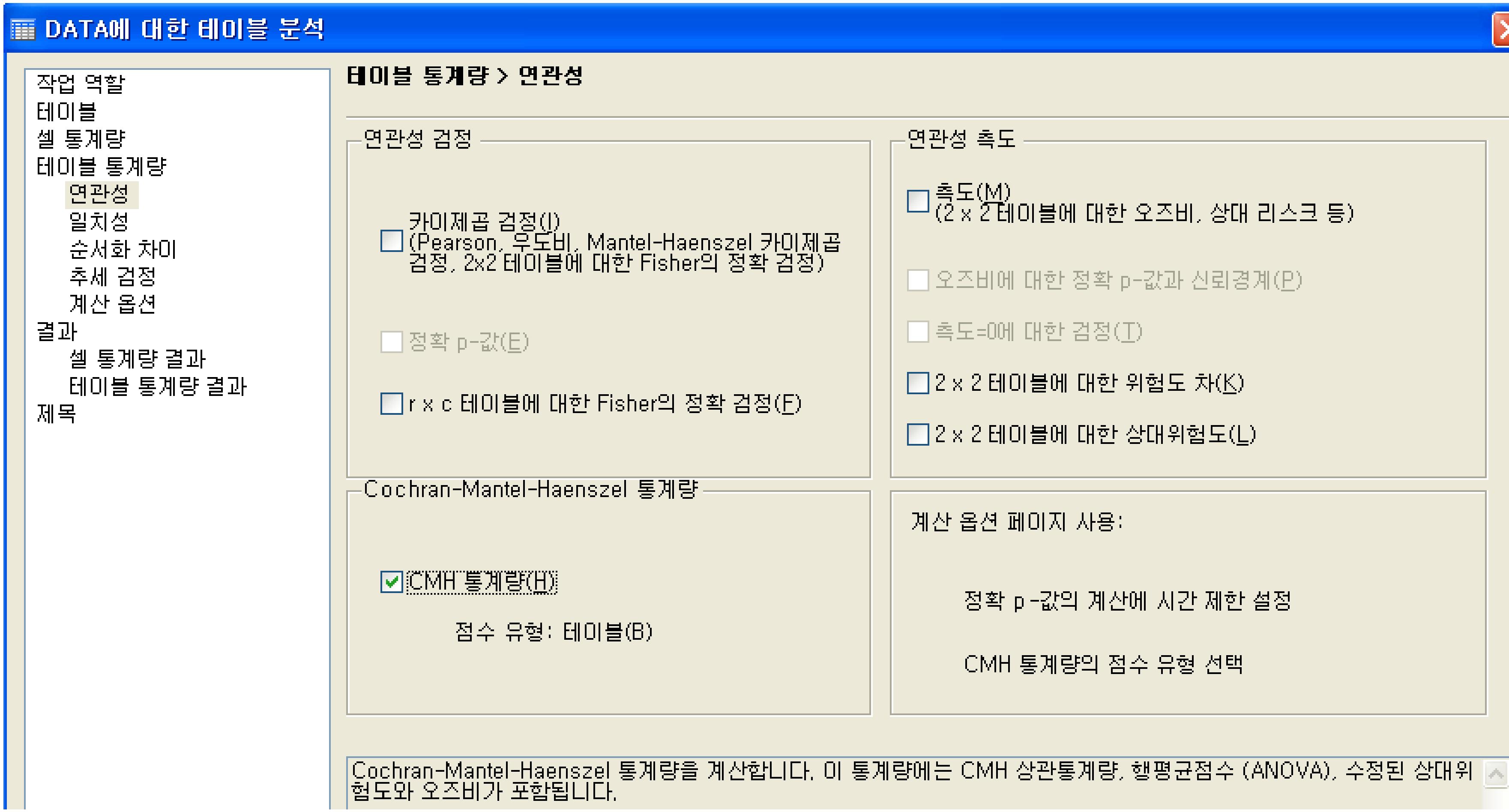
병원	치료제	결과		합계
		호전	비호전	
1 (종합병원)	기존	9	5	14
	신치료제	11	6	17
	계	20	11	31
2 (개인병원)	기존	7	5	12
	신치료제	8	3	11
	계	15	8	23
3 (개인의원)	기존	4	6	10
	신치료제	7	5	12
	계	11	11	22
4 (종합병원)	기존	18	11	29
	신치료제	26	4	30
	계	44	15	59
합계		90	45	135



	hospital	trt	result	obs
1	A	old	yes	9
2	A	old	no	5
3	A	new	yes	11
4	A	new	no	6
5	B	old	yes	7
6	B	old	no	5
7	B	new	yes	8
8	B	new	no	3
9	C	old	yes	4
10	C	old	no	6
11	C	new	yes	7
12	C	new	no	5
13	D	old	yes	18
14	D	old	no	11
15	D	new	yes	26
16	D	new	no	4


DATA에 대한 테이블 분석
작업 역할
테이블
셀 통계량
테이블 통계량
연관성
일치성
순서화 차이
추세 검정
계산 옵션
결과
셀 통계량 결과
테이블 통계량 결과
제목
작업 역할
할당할 변수(A):
이름
hospital
trt
result
obs
작업 역할(E):
빈도 (제한: 1개)
obs
그룹 분석변수
테이블 변수
result
trt
hospital


DATA에 대한 테이블 분석
작업 역할
테이블
셀 통계량
테이블 통계량
연관성
일치성
순서화 차이
추세 검정
계산 옵션
결과
셀 통계량 결과
테이블 통계량 결과
제목
테이블
테이블에 사용 가능한 변수(V):
미리 보기:
hospital
result
trt
생성될 테이블(T):
trt * result ; 테이블 분리 변수: hospital
<새로운 테이블을 정의하려면 여기를 클릭합니다.>


DATA에 대한 테이블 분석
작업 역할
테이블
셀 통계량
테이블 통계량
연관성
일치성
순서화 차이
추세 검정
계산 옵션
결과
셀 통계량 결과
테이블 통계량 결과
제목
테이블 통계량 > 연관성
연관성 검정
카이제곱 검정(I)
(Pearson, 우도비, Mantel-Haenszel 카이제곱 검정, 2x2 테이블에 대한 Fisher의 정확 검정)
정확 p-값(E)
r x c 테이블에 대한 Fisher의 정확 검정(F)
연관성 척도
척도(M)
(2 x 2 테이블에 대한 오즈비, 상대 리스크 등)
오즈비에 대한 정확 p-값과 신뢰경계(P)
척도=0에 대한 검정(T)
2 x 2 테이블에 대한 위험도 차(K)
2 x 2 테이블에 대한 상대위험도(L)
Cochran-Mantel-Haenszel 통계량
CMH 통계량(H)
점수 유형: 테이블(B)
계산 옵션 페이지 사용:
정확 p-값의 계산에 시간 제한 설정
CMH 통계량의 점수 유형 선택
Cochran-Mantel-Haenszel 통계량을 계산합니다. 이 통계량에는 CMH 상관통계량, 행평균점수 (ANOVA), 수정된 상대위험도와 오즈비가 포함됩니다.

테이블 trt * result에 대한 요약 통계량
제어 변수 : hospital

Cochran-Mantel-Haenszel 통계량 (테이블 스코어에 기반한)				
통계량	**대립가설**	**자유도**	**값**	**확률**
1	**영이 아닌 상관계수**	1	4.0391	0.0445
2	**행 평균 스코어 차이**	1	4.0391	0.0445
3	**일반 연관성**	1	4.0391	0.0445

[해석] 코크란-맨텔-핸젤통계량이 Q_CMH=4.0391이며, p값이 0.0445로 유의수준 5%보다 작으므로 귀무가설을 기각하게 된다. 따라서 병원 효과를 제어한 후 기존 치료제와 새로운 치료제의 효과 차이를 살펴본 결과 유의수준 5%수준에서 치료제의 효과는 차이가 있는 것으로 나타났다. 즉, 치료제와 환자의 호전도 사이에는 유의한 관련성이 존재한다고 생각된다.

상대 리스크의 추정값(행1/행2)				
연구 유형	방법	값	95% 신뢰한계	
사례대조연구	Mantel-Haenszel	0.4659	0.2213	0.9809
(오즈비)	로짓	0.4711	0.2200	1.0090
코호트	Mantel-Haenszel	0.6082	0.3724	0.9933
(칼럼1 리스크)	로짓	0.6460	0.3966	1.0524
코호트	Mantel-Haenszel	1.2796	1.0033	1.6320
(칼럼2 리스크)	로짓	1.2859	1.0131	1.6322

오즈비의 동질성에 대한 Breslow-Day 검정	
카이제곱	1.8947
자유도	3
Pr > ChiSq	0.5946

- SAS program

```
Data test_ex3;
  input  hospital $ trt $  result $  obs @@;
  cards;
  A old yes  9  A old no  5  A new yes 11 A new no 6
  B old yes  7  B old no  5  B new yes 8  B new no 3
  C old yes  4  C old no  6  C new yes 7  C new no 5
  D old yes 18 D old no 11 D new yes 26 D new no 4
;
Proc freq;
 Weight obs;
 Tables hospital*trt*result/CMH;
Run;
```

Summary Statistics for trt by result
Controlling for hospital

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	4.0391	0.0445
2	Row Mean Scores Differ	1	4.0391	0.0445
3	General Association	1	4.0391	0.0445

실험군 : 1
대조군 : 2

Estimates of the Common Relative Risk (Row1/Row2)

오즈비

Type of Study	Method	Value	95% Confidence Limits	
Case-Control	Mantel-Haenszel	0.4659	0.2213	0.9809
(Odds Ratio)	Logit	0.4711	0.2200	1.0090
Cohort	Mantel-Haenszel	0.6082	0.3724	0.9933
(Col1 Risk)	Logit	0.6460	0.3966	1.0524
Cohort	Mantel-Haenszel	1.2796	1.0033	1.6320
(Col2 Risk)	Logit	1.2859	1.0131	1.6322

상대위험률
(발병기준 2)

Breslow-Day Test for
Homogeneity of the Odds Ratios

Chi-Square	1.8947
DF	3
Pr > ChiSq	0.5946

Total Sample Size = 135

3.연관성측도

연관성 측도의 이론 개념

- 행과 열의 범주가 각각 2인 2X2 분할표의 경우에 널리 이용되는 연관성 측도
- 역학 연구 분야의 코호트 연구나 사례-대조 연구에서의 중요한 모수

		Exposure		
		Present E	Absent $\bar{E}$	Total
Disease	Present D	a	b	n_1
	Absent $\bar{D}$	c	d	n_2
	Total	m_1	m_2	**n**

$$P_1 = \Pr(D \mid E),\ \ 1 - P_1 = \Pr(\bar{D} \mid E),\ \ P_2 = \Pr(D \mid \bar{E}),\ \ 1 - P_2 = \Pr(\bar{D} \mid \bar{E})$$

$$P_1' = \Pr(E \mid D),\ \ 1 - P_1' = \Pr(\bar{E} \mid D),\ \ P_2' = \Pr(E \mid \bar{D}),\ \ 1 - P_2' = \Pr(\bar{E} \mid \bar{D})$$

- **Relative Risk (RR, 상대 위험율)**
  - **표현식과 추정량**

$$RR = \frac{\text{실험군에서의 위험률}}{\text{대조군에서의 위험률}} = \frac{\Pr(D \mid E)}{\Pr(D \mid \overline{E})} = \frac{P_1}{P_2}$$

$$R\hat{R} = \frac{a / m_1}{b / m_2}$$

  - 코호트 연구나 사례군과 대조군이 미리 정해진 후 그 결과를 관찰하는 임상시험 연구에서 계산 가능
  - 사례-대조 연구에서는 계산 불가능

- **RR 의 표본분포**
  - $\hat{RR}$ 의 표본분포 특성
    - 매우 큰 표본에서는 RR은 근사정규분포
    - 역학연구의 표본크기 수준에서는 정규분포 형태가 아니며, 양의 왜도를 나타내는 분포
      - 자연로그 변환 $\ln(\hat{RR})$ 을 통해 보다 대칭적인 분포 형태를 취하게 되며, 정규분포로 근사
  - 추정량 : 코호트 연구에서 계산 가능하므로 $\hat{RR} = p_1/p_2$
  - $\ln(\hat{RR})$ 의 분산과 추정량

$$Var[\ln(\hat{RR})] = Var[\ln(p_1) - \ln(p_2)] = Var[\ln(p_1)] + Var[\ln(p_2)]$$

$$\hat{V}ar[\ln(\hat{RR})] = \left(\frac{1}{m_1}\right)\frac{1-p_1}{p_1} + \left(\frac{1}{m_2}\right)\frac{1-p_2}{p_2} \quad \text{since} \quad Var[\ln(p)] = \left(\frac{1}{n}\right)\frac{1-p}{p}$$

  - 신뢰구간은 OR과 동일한 특성을 보임

– 오즈비(OR : Odds Ratio)

$$OR = \frac{\text{실험군의 오즈}}{\text{대조군의 오즈}} \quad , \quad \text{오즈} = \frac{\text{사건이 발생한 확률}}{\text{사건이 발생하지 않은 확률}}$$

- 사례-대조 연구 (case-control study)에 주로 이용

- **Odds Ratio (OR, 오즈비)**
  - **표현식 : 코호트 연구**

$$\mathrm{OR} = \frac{\text{실험군에서의 오즈비}}{\text{대조군에서의 오즈비}} = \frac{O_1}{O_2} = \frac{\Pr(D \mid E)/\Pr(\bar{D} \mid E)}{\Pr(D \mid \bar{E})/\Pr(\bar{D} \mid \bar{E})} = \frac{P_1/(1-P_1)}{P_2/(1-P_2)}$$

  - **표현식 : 사례-대조연구**

$$\mathrm{OR} = \frac{\text{실험군에서의 오즈비}}{\text{대조군에서의 오즈비}} = \frac{O_1}{O_2} = \frac{\Pr(E \mid D)/\Pr(\bar{E} \mid D)}{\Pr(E \mid \bar{D})/\Pr(\bar{E} \mid \bar{D})} = \frac{P_1'/(1-P_1')}{P_2'/(1-P_2')}$$

  - **추정량 : the cross product ratio**

$$\hat{\mathrm{OR}} = \frac{\left(a/m_1\right)\big/\left(c/m_1\right)}{\left(b/m_2\right)\big/\left(d/m_2\right)} \; or \; \frac{\left(a/n_1\right)\big/\left(b/n_1\right)}{\left(c/n_2\right)\big/\left(d/n_2\right)} = \frac{ad}{bc}$$

- **OR 의 표본분포**
  - OR의 특성
    - OR의 범위 $0 \le OR \le \infty$
    - $OR = 1$ 의 의미 : 위험율이 노출그룹과 비노출그룹, 혹은 사례군과 대조군의 오즈비가 동일함을 의미
    - $O\hat{R}$ 의 표본 분포 특성
      - 비정규분포, 꼬리가 큰 값으로 긴 강한 양의 왜도를 갖는 분포
      - 자연로그변환 $\ln(O\hat{R})$ 을 통한 분포 특성 개선 : 보다 근사정규분포
      - 신뢰구간(표준정규분포 이용)은 비대칭
  - $\ln(O\hat{R})$ 의 분산과 추정량

$$\mathrm{Var}[\ln(O\hat{R})] \approx \left(\frac{1}{n_1}\right)\frac{1}{P_1'(1-P_1')} + \left(\frac{1}{n_2}\right)\frac{1}{P_2'(1-P_2')}$$

$$\hat{V}ar[\ln(O\hat{R})] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

  - $H_0 : OR = 1 \text{ vs } H_1 : OR \neq 1$ 에 대한 검정 : 2X2 카이제곱검정

연관성 측도 : 예제 / 상대위험률

- 아스피린이 심장병의 위험을 줄이는 데 효과가 있는 가에 대한 연구 (Dawson-Saunders & Trapp, 1994 : 이재원 외 인용 p. 40)

	심근경색 발병		전체
	유	무	
아스피린	139	10,898	11,037
플라시보	239	10,795	11,034

  - 상대위험률(RR : relative risk)

$$RR = \frac{\text{실험군의 위험률}}{\text{대조군의 위험률}} = \frac{139 \ / 11037}{239 \ / 11034} = .581$$

    - 해석 : 상대위험률이 “1”보다 작으므로 아스피린을 복용할 때 그렇지 않은 경우에 비해 심근경색을 일으킬 위험이 상대적으로 낮다는 의미(참조 : “1”이면 차이가 없다는 의미임)

연관성 측도 : 예제/오즈비

- 심장발작을 일으킨 환자와 그렇지 않은 환자를 대상으로 과거 약물 남용 경험을 조사한 사례-대조 연구(Dawson-Saunders & Trapp, 1994 : 이재원 외 인용 p. 41)

약물 남용 경험	심장 발작	
	유	무
유	73	18
무	141	196
전체	214	214

- 실험군 odds

$$= \frac{73\ /\ 214}{141\ \ /\ 214} = 0.518$$

- 대조군 odds

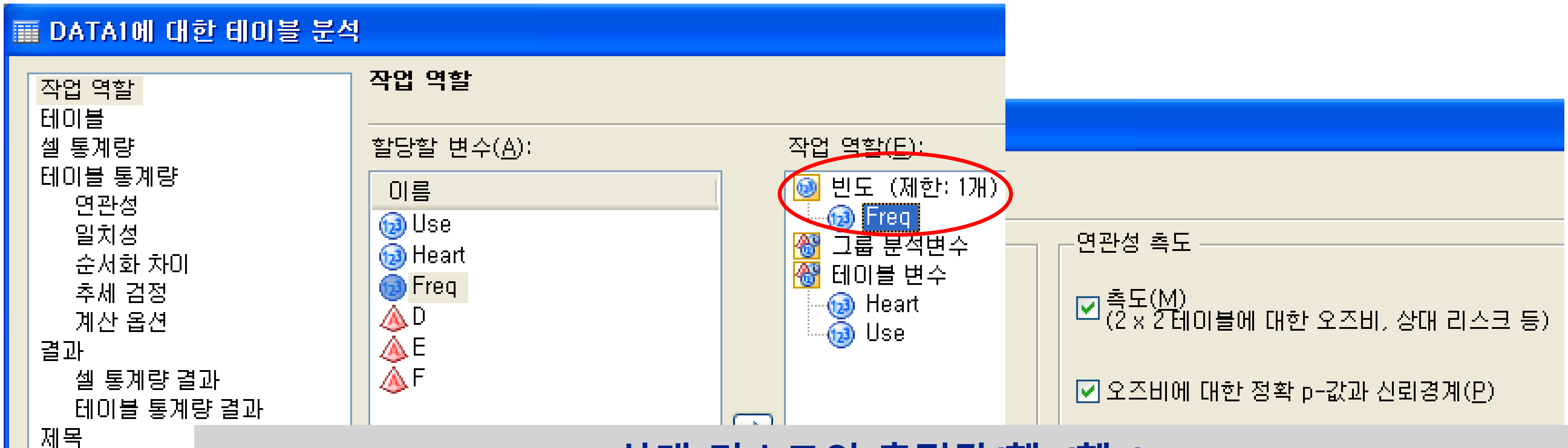
$$= \frac{18\ /\ 214}{196\ \ /\ 214} = 0.092$$

– 오즈비(OR : Odds Ratio)

$$OR = \frac{\text{실험군의}\quad \text{오즈}}{\text{대조군의}\quad \text{오즈}} = \frac{0.518}{0.092} = 5.64$$

- 해석 : 실험군에서 약물남용 환자들이 대조군보다 5.64배 많다는 의미(참조 : “1”이면 연관성이 없다는 의미임)

연관성 측도 : SAS 결과 및 해석



상대 리스크의 추정값(행1/행2)

연구 유형	값	95% 신뢰한계	
사례대조연구 (오즈비)	5.6375	3.2223	9.8631
코호트 (칼럼1 리스크)	1.9173	1.6305	2.2546
코호트 (칼럼2 리스크)	0.3401	0.2227	0.5195

➢ 해석 : 오즈비가 5.64로 실험군에서 약물남용 환자가 상대적으로 많다는 의미이며, 95% 신뢰구간이 "1"을 포함하지 않으므로 약물남용이 심장발작의 위험률을 높인다고 분석가능

연관성 척도 : 활용 방안(1)

- 예 : KB 은행과 신한은행의 유사 경쟁상품(펀드 가입유무를 고려한) 경쟁력 비교

	경쟁 상품 가입여부		전체
	유	무	
KB	a	b	a+b
신한	c	d	c+d

– 상대경쟁력(PRC : Power of Relative Competition)

$$\text{상대경쟁력} = \frac{\text{비교은행의 가입률}}{\text{경쟁은행의 가입률}} = \frac{a/(a+b)}{c/(c+d)}$$

연관성 측도 : 활용방안(2)

- 예: KB은행의 만족도가 높은 그룹과 낮은 그룹을 대상으로 특정상품 가입유무를 조사한 자료를 가정

특정상품 가입유무	만족도	
	높음	낮음
유	a	b
무	c	d
전체	a+c	b+d

– 만족도가 높은 집단의 상대충성도 $= \dfrac{a/(a+c)}{c/(a+c)}$

– 만족도가 낮은 집단의 상대충성도 $= \dfrac{b/(b+d)}{d/(b+d)}$

– 상대충성도 지수 (Relative Royalty Index : RRI)

$$RRI = \frac{\text{만족도가 높은 집단의 상대충성도}}{\text{만족도가 낮은 집단의 상대충성도}}$$

4.진단법의 평가

진단법의 평가 : 민감도와 특이도

- 민감도(sensitivity) : 질병이 있는 사람을 양성으로 판정하는 비율

$$\text{민감도} = \frac{TP}{TP + FN}$$

- 특이도(specificity) : 질병이 없는 사람을 음성으로 판정하는 비율

$$\text{특이도} = \frac{TN}{FP + TN}$$

- 해석 :
  - 민감도가 높다면, 진단은 질병에 걸린 사람을 음성으로 잘못 진단하는 위음성율(false-negative rate)이 낮다는 의미
  - 특이도가 높다면, 질병에 걸리지 않은 사람을 양성으로 잘못진단하는 위양성율(false-positive rate)이 낮다는 의미

민감도와 특이도 : 예제

		결핵		계
		D+	D-	
X-ray	T+	22	51	73
	T-	8	1739	1747
계		30	1790	1820

(자료 출처 : 이재원 외 p. 60 인용)

– $\text{민감도} = \frac{TP}{TP + FN} = \frac{22}{30} = 0.73$

– $\text{특이도} = \frac{TN}{FP + TN} = \frac{1739}{1790} = 0.97$

- 해석 :
  - 결핵에 대한 진단방법은 민감도는 0.73이며, 특이도는 0.97임.

양성예측도와 음성예측도의 추정

- 기본개념
  - 진단을 통해 양성으로 판정된 경우 정말 질병에 걸렸을 확률 혹은 음성으로 판정된 경우 정말 질병에 걸리지 않았을 확률의 추정 문제도 중요
  - 베이즈 정리를 이용하여 확률 계산
  - 양성예측도(predicted value of positive test) :

$$P(D+|T+) = \frac{P(T+\cap D+)}{P(T+)} = \frac{P(T+|D+)\,P(D+)}{P(T+|D+)\,P(D+) + P(T+|D-)\,P(D-)}$$

$$= \frac{\text{민감도} \times \text{유병률}}{\text{민감도} \times \text{유병률} + (1-\text{특이도}) \times (1-\text{유병률})}$$

  - 음성예측도(predicted value of negative test)

$$P(D-|T-) = \frac{P(T-\cap D-)}{P(T-)} = \frac{P(T-|D-)\,P(D-)}{P(T-|D-)\,P(D-) + P(T-|D+)\,P(D+)}$$

$$= \frac{\text{특이도} \times (1-\text{유병률})}{\text{특이도} \times (1-\text{유병률}) + (1-\text{민감도}) \times \text{유병률}}$$

양성예측도와 음성예측도의 추정 : 예제

- 예제 자료 : ppt p. 72 계속 (결핵 자료)
  - 민감도=0.73, 특이도=0.97
  - 유병률 : 인구 10만명당 9.3명으로 결핵 발생(1987년 결과)
    (유병률=0.000093)
- $$\text{양성예측도} = \frac{\text{민감도} \times \text{유병률}}{\text{민감도} \times \text{유병률} + (1-\text{특이도}) \times (1-\text{유병률})}$$
  $$= \frac{0.73 \times 0.000093}{0.73 \times 0.000093 + (1-0.97) \times (1-0.000093)} = 0.002258$$
- $$\text{음성예측도} = \frac{\text{특이도} \times (1-\text{유병률})}{\text{특이도} \times (1-\text{유병률}) + (1-\text{민감도}) \times \text{유병률}}$$
  $$= \frac{0.97 \times (1-0.000093)}{0.97 \times (1-0.000093) + (1-0.73) \times 0.000093} = 0.999974$$

- 해석 :
  - X-ray 진단에 의해 양성으로 판정된 사람 중 0.23%만이 실제 결핵환자로 추정되며, 음성으로 판정된 사람 중 99.997%는 결핵환자가 아니라고 판단됨.